Analyzing related raw data files through dataflows

Vitor Silva, Daniel de Oliveira, Patrick Valduriez and Marta Mattoso

COPPE - Federal University of Rio de Janeiro

COPPE
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia
UFRJ

Distributed & Parallel Data Mngmt

- **Scientific Workflow Management Systems**
- **Data parallelism**
- **Scientific data analysis through Provenance**
- **Data & task parallelism in Clusters, Clouds**

Hoscar's objectives – 2012
Group 2: Scientific Data Management

Processing of very large datasets.

- This topic deals with the management of very large datasets that are manipulated (accessed and produced) by data-centric scientific workflows in HPC environments.
- Addressing the very scale of the datasets requires new scalable parallel data management techniques as well as scalable data-aware scheduling strategies.
- Furthermore, getting data in and out HPC environments from the scientists' own environment is a major challenge.
- Finally, it is important to provide support for data provenance, a key function that records critical metadata about experiments to help scientists understanding results or reusing some workflow parts.

Current results:

- New CS techniques on these 4 issues
- joint publications among CS-BR & CS-FR
- validations with scientific applications from BR partners

Challenges for HPC4E:

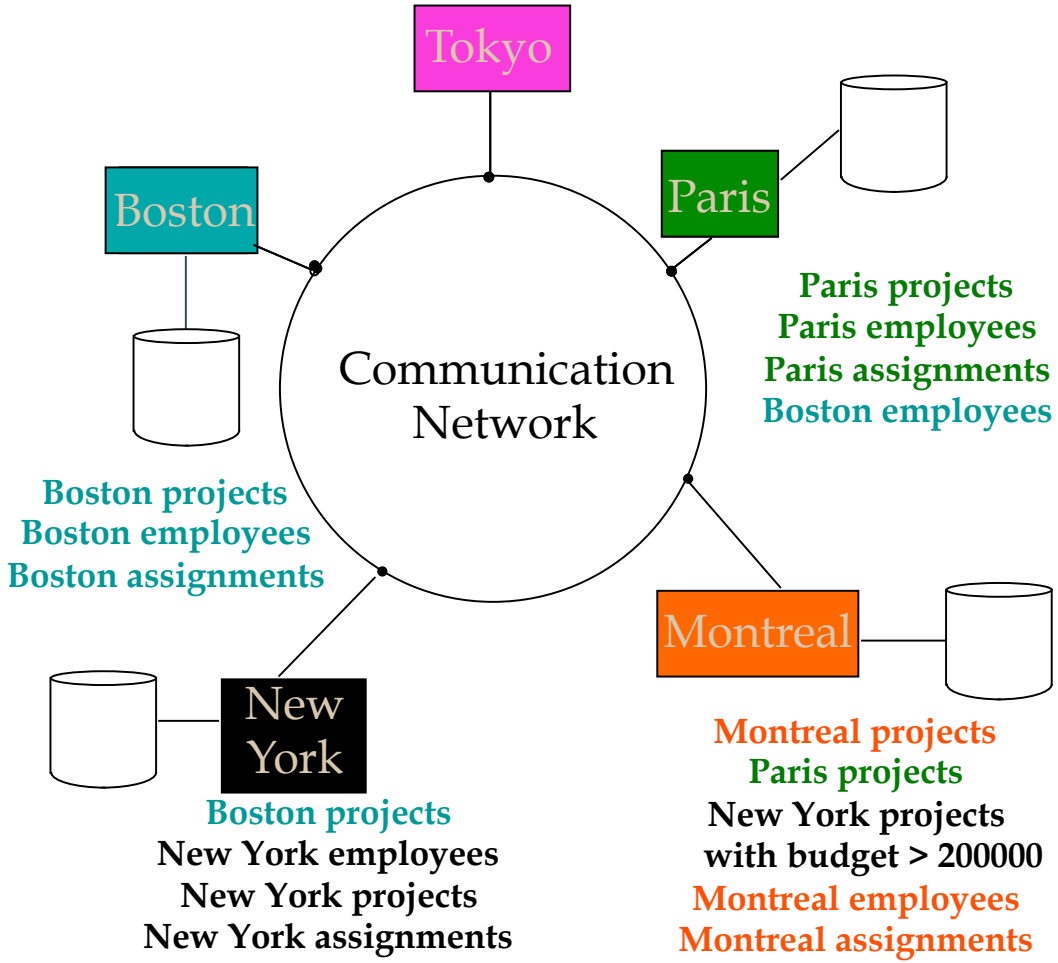
- Traversal actions between groups

http://www-sop.inria.fr/hoscar/pmwiki-2.2.40/pmwiki.php/Main/Objectives

What Data Management can do:

data integration, replication, partition, indexing, concurrency, consistency, fault tolerance

```
SELECT ENAME,SAL
FROM   EMP,ASG,PAY
WHERE  DUR > 12
AND    EMP.ENO = ASG.ENO
AND    PAY.TITLE = EMP.TITLE
```



From M. T. Özsu & P. Valduriez "Principles of Distributed Databases", book by Springer

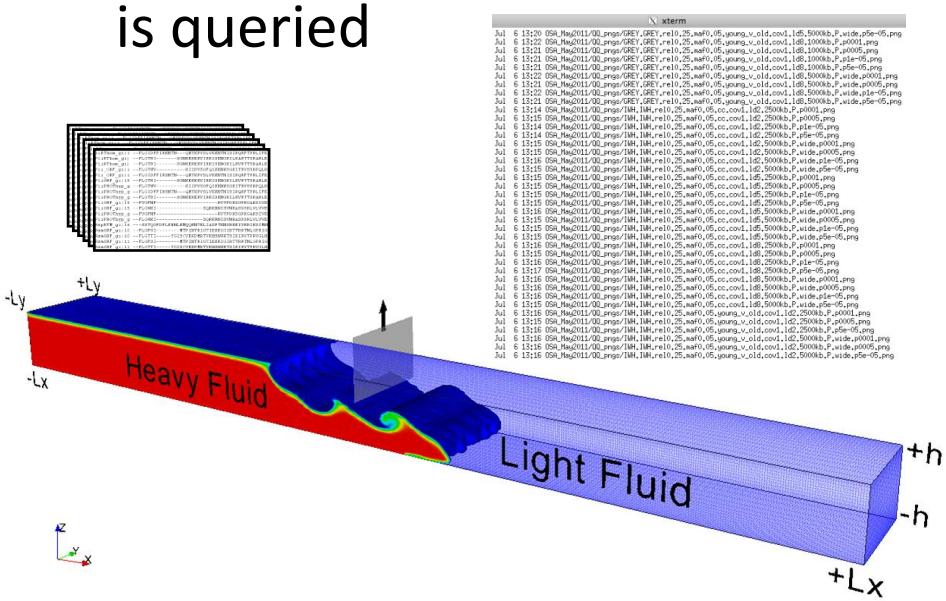
Why Data Management differs

Business data

- easy to understand
- text format
- all manipulation in SQL
- most of data stored is traversed for queries

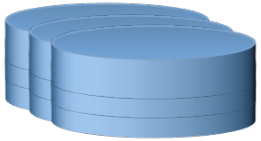
Scientific data

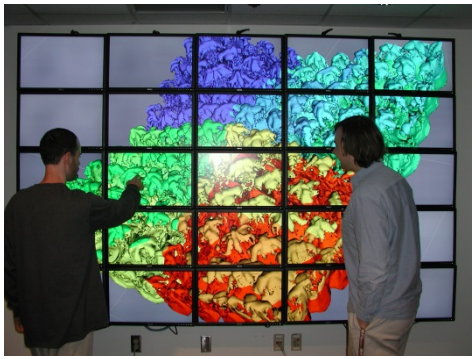
- complex maths / domain
- binary shared formats
- specific programs access
- only a small fraction of data is queried



Scientific Data Management
- let your data BE managed !!!

- data transformations – *ad-hoc*
- files generated independently
- parallel processing unaware of data-flow
- analysts need to manually manage the larger life cycle of big data flow analysis





xterm

```
Jul  6 13:20 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld5.5000kb.P.wide.p5e-05.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.1000kb.P.p0001.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.1000kb.P.p0005.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.1000kb.P.p1e-05.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.1000kb.P.p5e-05.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.5000kb.P.wide.p0001.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.5000kb.P.wide.p0005.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.5000kb.P.wide.p1e-05.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY.rel0.25.maf0.05.young_v_old.cov1.ld8.5000kb.P.wide.p5e-05.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.2500kb.P.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.2500kb.P.p0005.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.2500kb.P.p1e-05.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.2500kb.P.p5e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.5000kb.P.wide.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.5000kb.P.wide.p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.5000kb.P.wide.p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld2.5000kb.P.wide.p5e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.2500kb.P.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.2500kb.P.p0005.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.2500kb.P.p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.2500kb.P.p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.5000kb.P.wide.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.5000kb.P.wide.p0005.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.5000kb.P.wide.p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld5.5000kb.P.wide.p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.2500kb.P.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.2500kb.P.p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.2500kb.P.p1e-05.png
Jul  6 13:17 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.2500kb.P.p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.5000kb.P.wide.p0001.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.5000kb.P.wide.p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.5000kb.P.wide.p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.cc.cov1.ld8.5000kb.P.wide.p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.2500kb.P.p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.2500kb.P.p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.2500kb.P.p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.5000kb.P.wide.p0001.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.5000kb.P.wide.p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH.rel0.25.maf0.05.young_v_old.cov1.ld2.5000kb.P.wide.p5e-05.png
```

BLOG@CACM

Data Science Workflow: Overview and Challenges, by Philip Guo

Raw data file analysis - limitations

- Load into a DB
  - time consuming, specific data structures (e.g. graphs, matrix, images, visualization)
- Build an access structure and library to a specific format
  - partial DB load, but too specific
- Build an access structure and library to a generic format loading the database as the queries are formulated
  - partial DB load, indexing on the fly, as needed
- no dataflow, no runtime queries

Putting the human in the loop

"In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding."

Exploring the inherent technical challenges in realizing the potential of Big Data.

BY H.V. JAGADISH, JOHANNES GEHRKE,
ALEXANDROS LABRINIDIS, YANNIS PAPAKONSTANTINOU,
JIGNESH M. PATEL, RAGHU RAMAKRISHNAN,
AND CYRUS SHAHABI

Big Data and Its Technical Challenges

Our solution

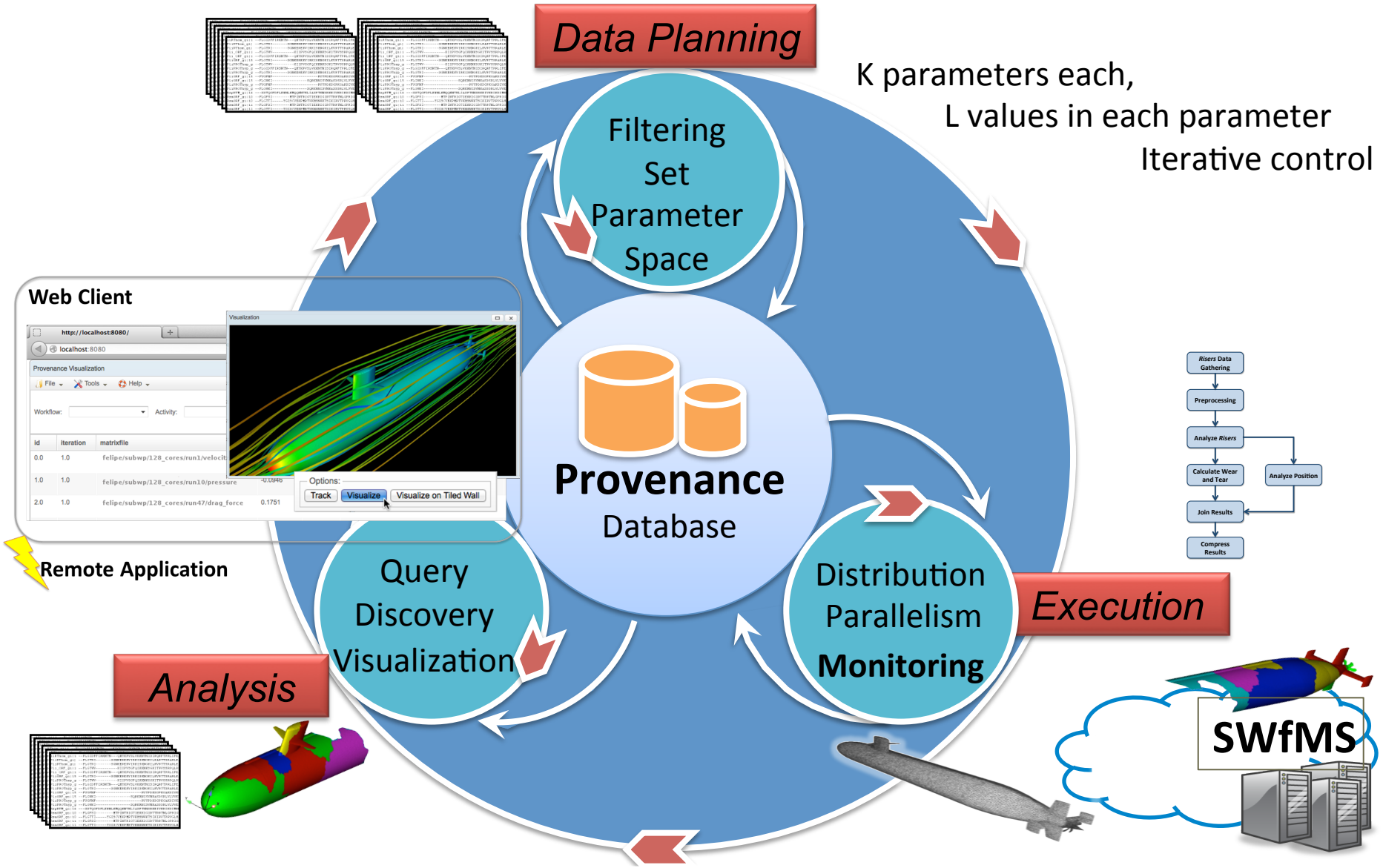
- Load selected data while it is being generated (e.g. quantities of interest)
- Runtime query (HIL)
- Relates several raw data files (Dataflow aware)
- Takes advantage of
  - a relational algebraic approach
  - provenance data
  - parallel processing

Important data is extracted: e.g. energy, velocity, pressure, …

- Indexing on the fly
- Querying on pre-selected data to restrict raw data file analysis
- Multiple files are related
- Queries can be dataflow aware
- Few selected data loading before querying
- Limitations
  - when relevant data is not extracted
  - can be complemented by related work

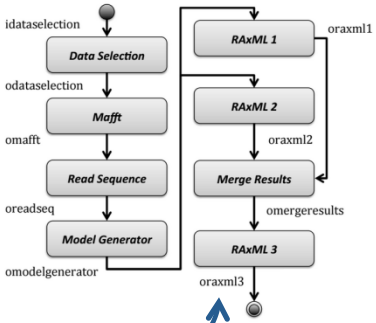
Workflow Life Cycle

during execution- user steering



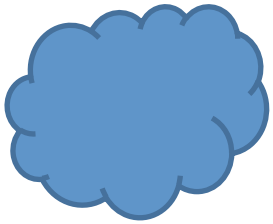
Using a DBMS to manage parameter sweep, task scheduling & provenance

2. Scientific Application Modeling using Data-centric Workflows



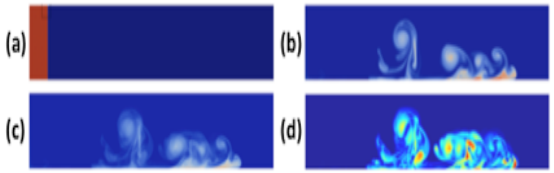
1. Scientific Composition

3. Managed by a **SWfMS** and executed in HPC environments

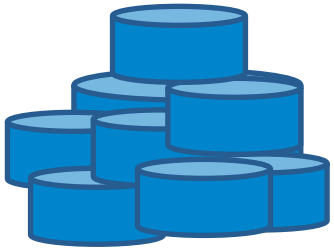


Database

5. Data Analyses



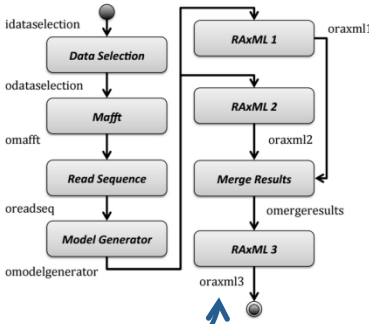
4. **Metadata** from large volumes of data manipulated



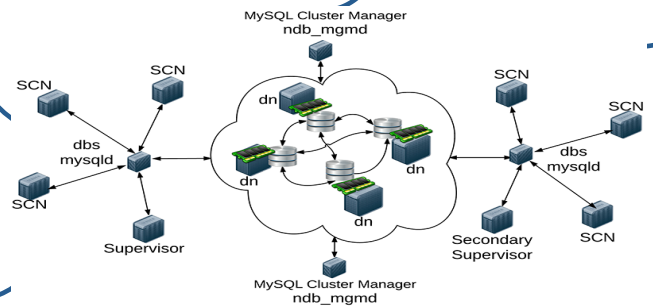
- Petabytes
- Exabytes

Using a **D-**DBMS to manage parameter sweep task, scheduling & provenance

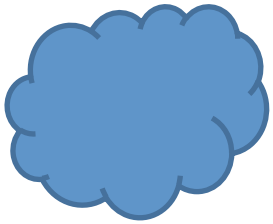
2. Scientific Application Modeling using Data-centric Workflows



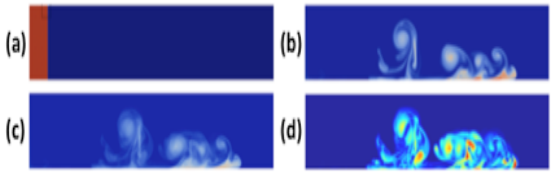
1. Scientific Composition



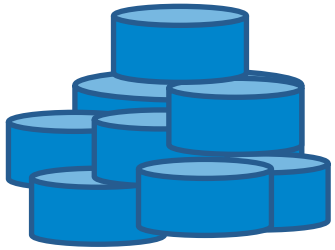
3. Managed by a **SWfMS** and executed in HPC environments



5. Data Analyses

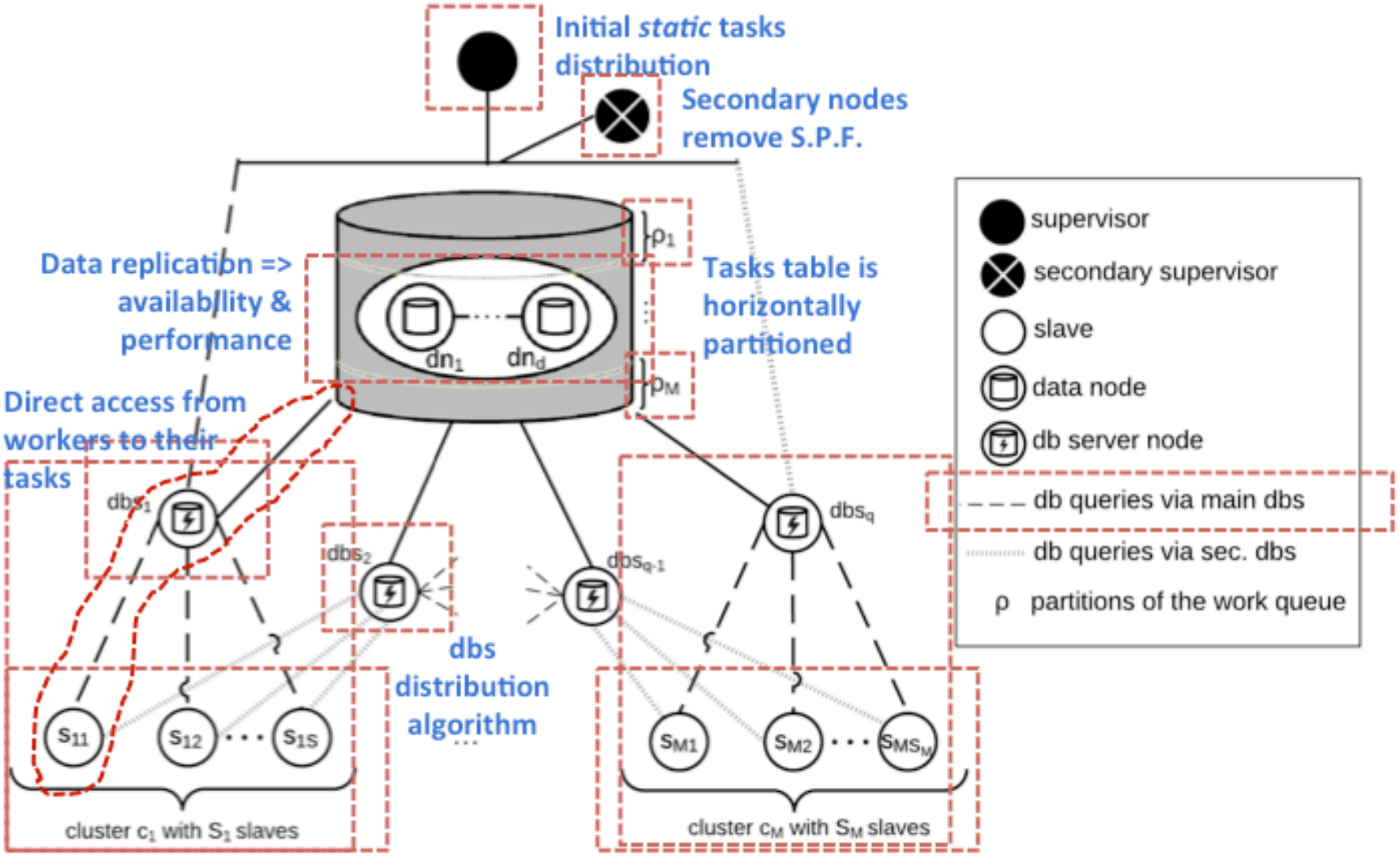


4. **Metadata** from large volumes of data manipulated



- Petabytes
- Exabytes

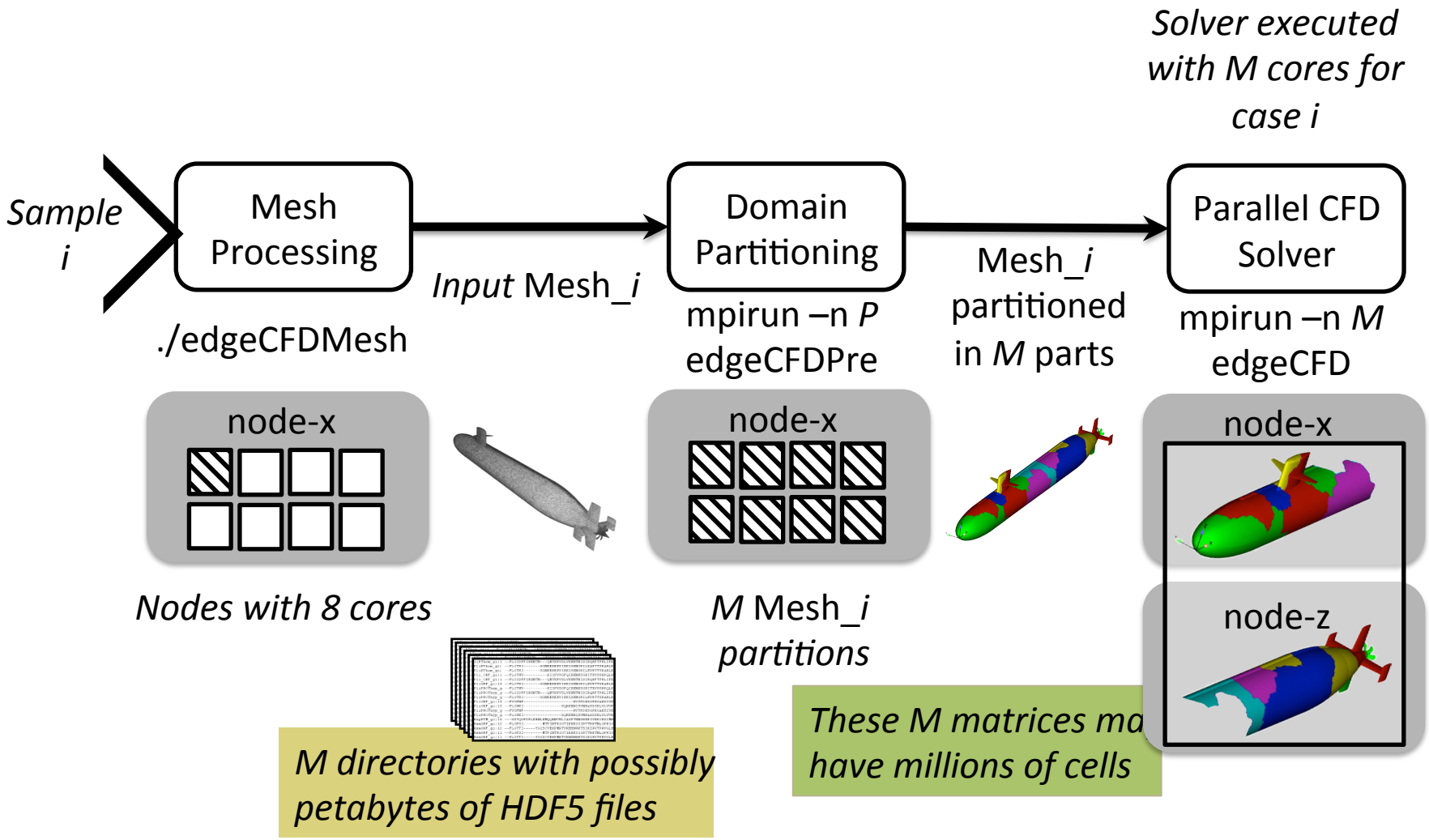
Controlling Tasks Scheduling Relying on a DDBMS



R. Souza, V. Silva, D. de Oliveira, P. Valduriez, A. Lima, M. Mattoso. Parallel Execution of Workflows Driven by a Distributed Database Management System **IEEE/ACM SuperComputing Conference** *(SC15)*, 2015

If resources are disconnected-

How to associate initial files to parameters and images



If resources are disconnected-

How to query initial files, parameters and images

Several files, matrices, parameters, results just for exploration ***i***

Typical runtime queries:

- solver is converging?
- input tolerance are satisfied ?
- partial visualiz indicates good sol?

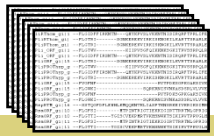
Solver executed with 16 cores for case i

Parallel CFD Solver

mpirun –n *M* edgeCFD

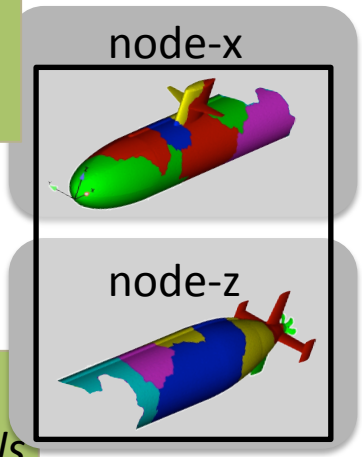
Nodes with 8 cores

16 Meshi *partitions*



Several directories with petabytes of HDF5 files

These M matrices have millions of cells



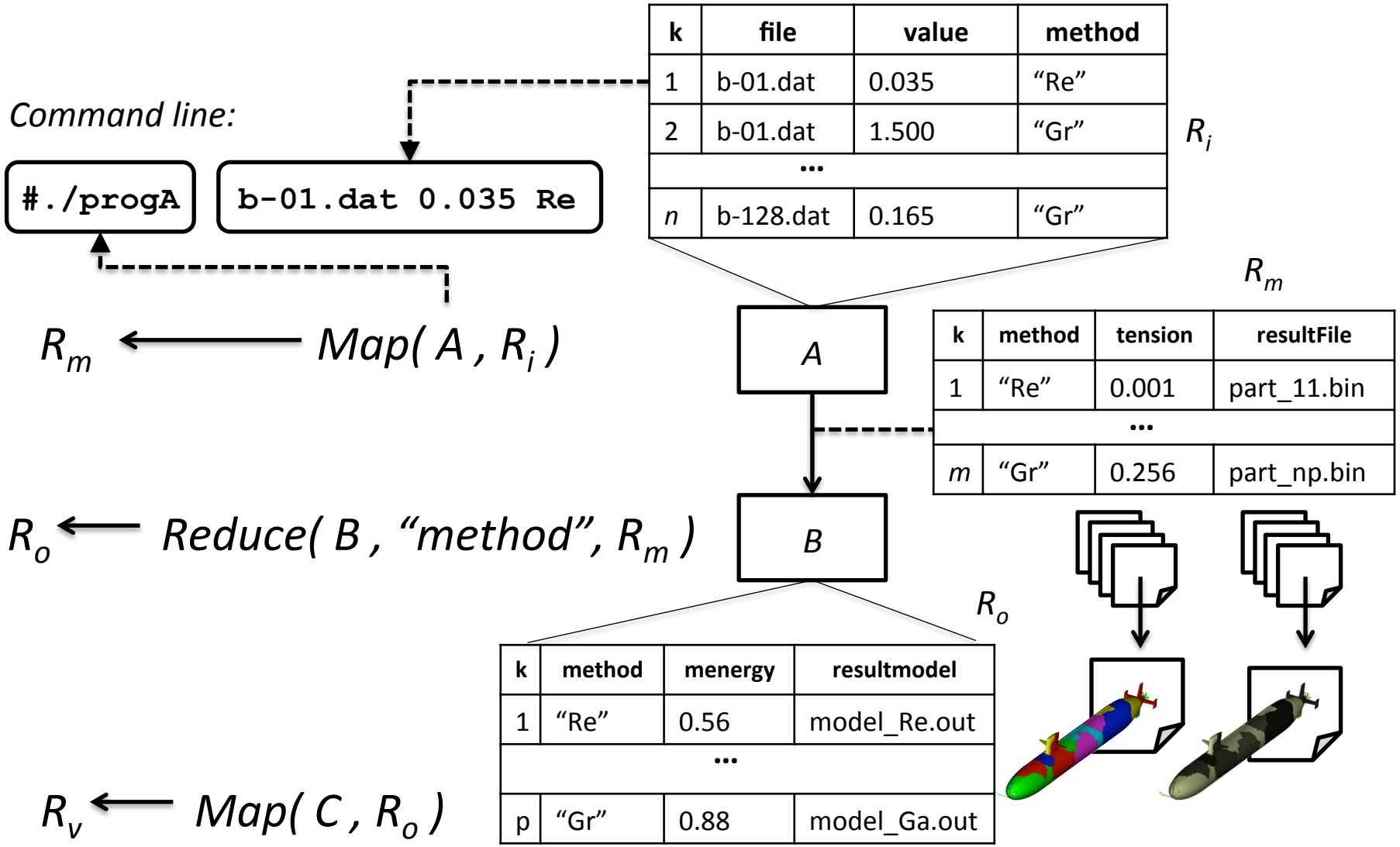
SWfMS & Provenance

- Scientific Workflow Management Systems (SWfMS)
- Efficient execution of scientific workflows
- Tracing the execution through provenance
- **Provenance** data (W3C PROV working group)
  - to enable scientific discovery
  - reproducibility,
  - result interpretation, and
  - problem diagnosis in scientific experiments

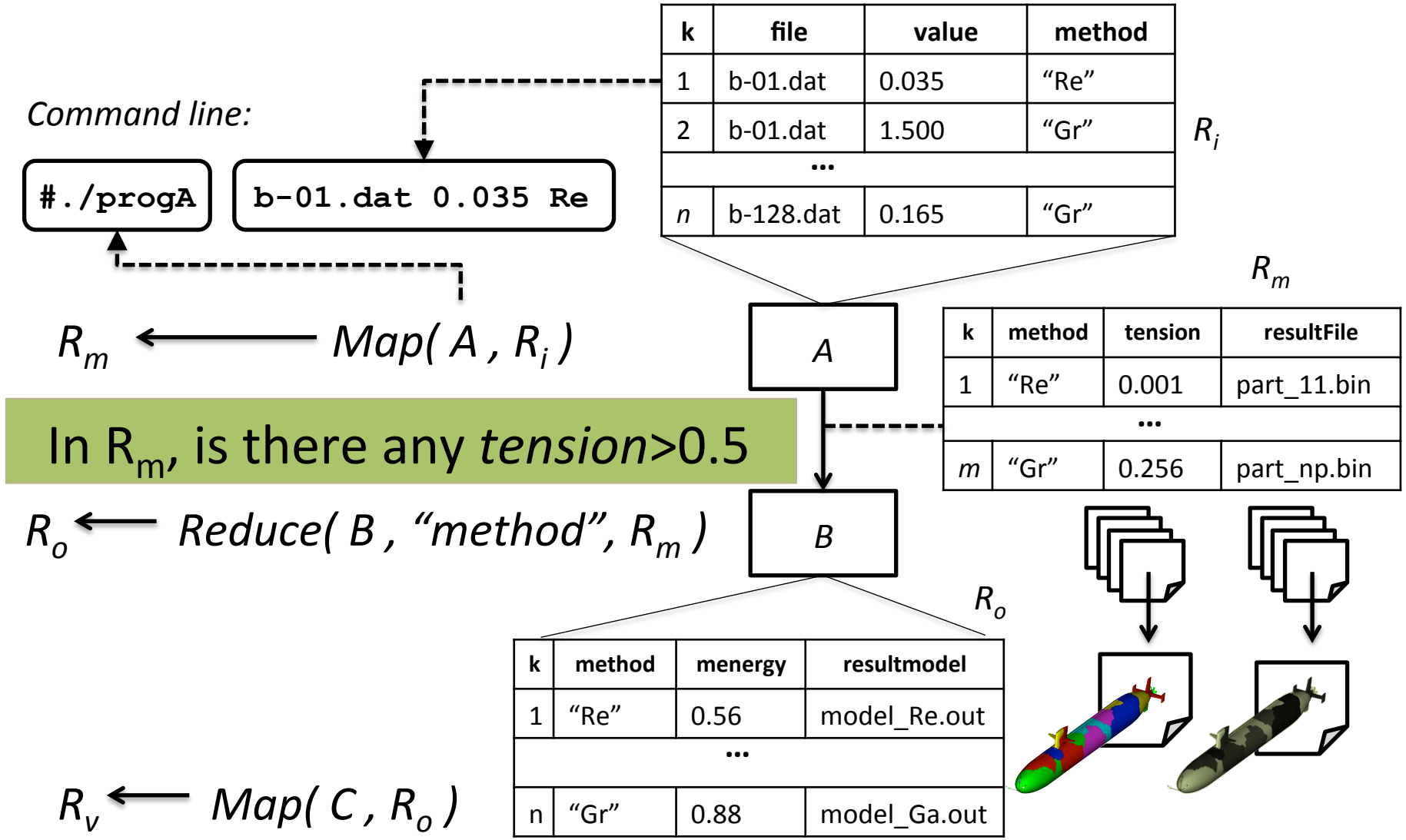
Enabling technologies

- Scientific Workflow Management Systems
- Data parallelism similar to MapReduce
- Provenance data analysis
- Chiron's Dataflow Algebraic Approach
  - non intrusive wrt parallel numerical solution
  - online data analysis
  - convergence tracking
  - visualization of partial results
  - dynamic interference on loop parameters

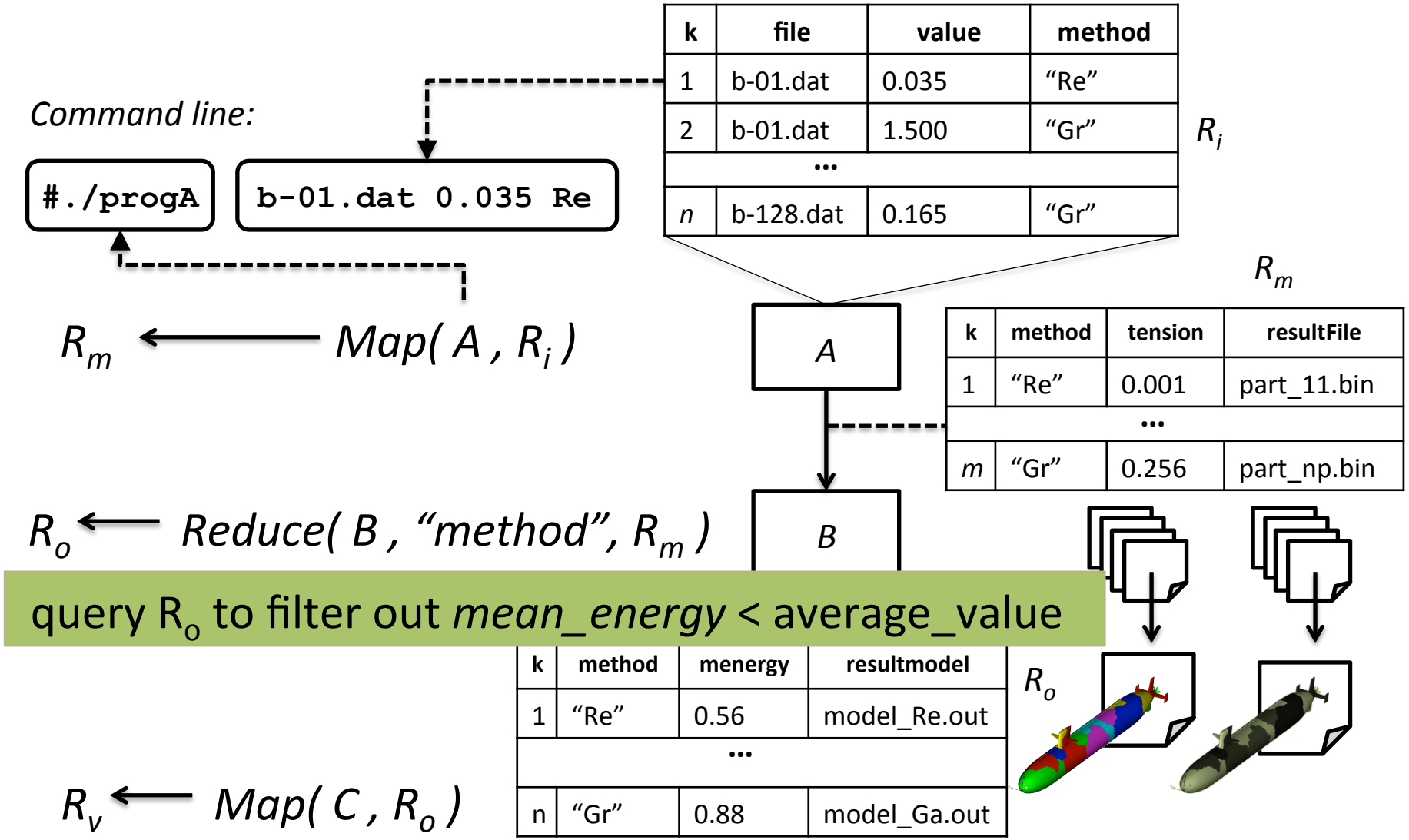
Dataflow Algebraic Workflow Engine



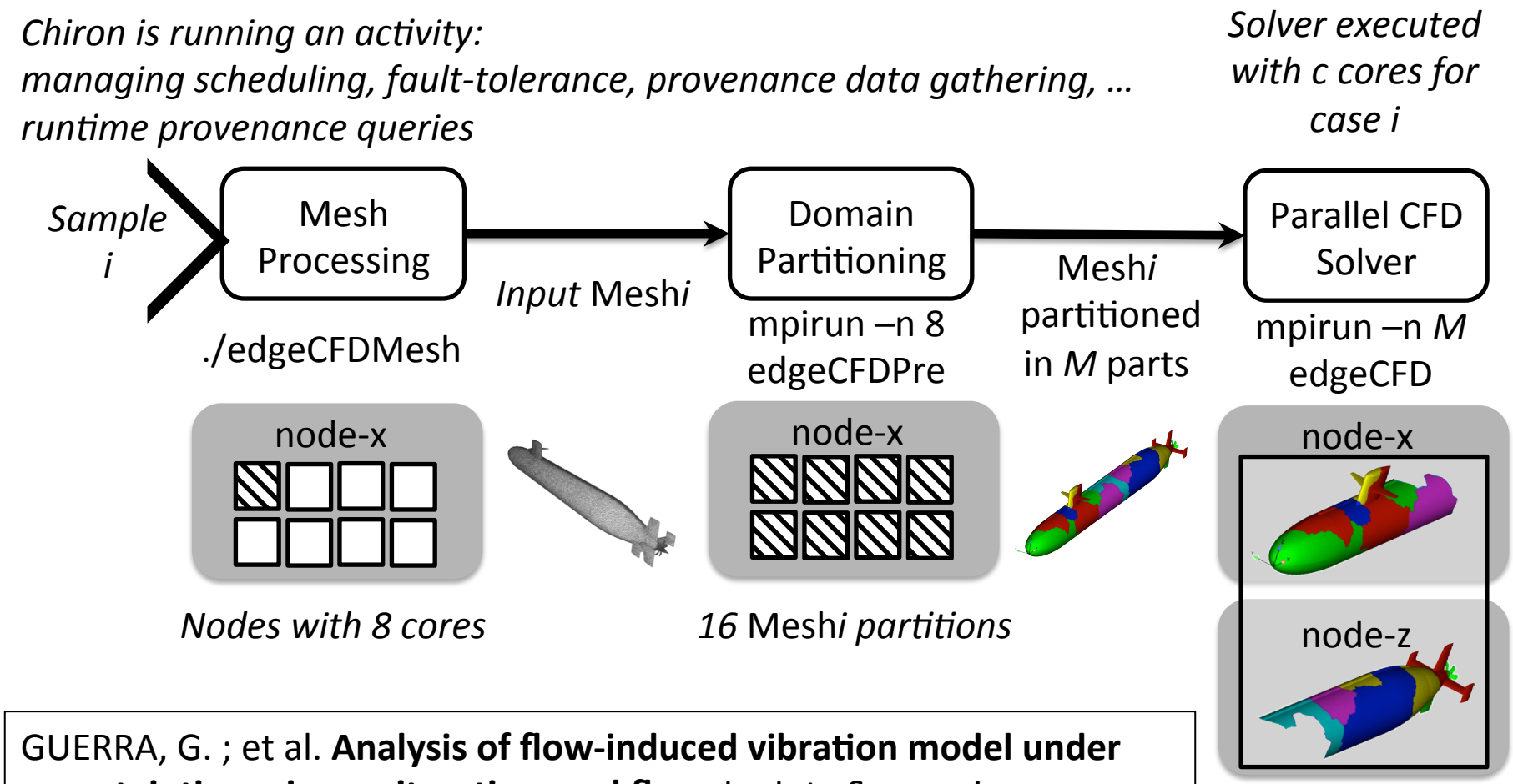
As the workflow executes … user steering (HIL)



As the workflow executes … user steering (HIL)

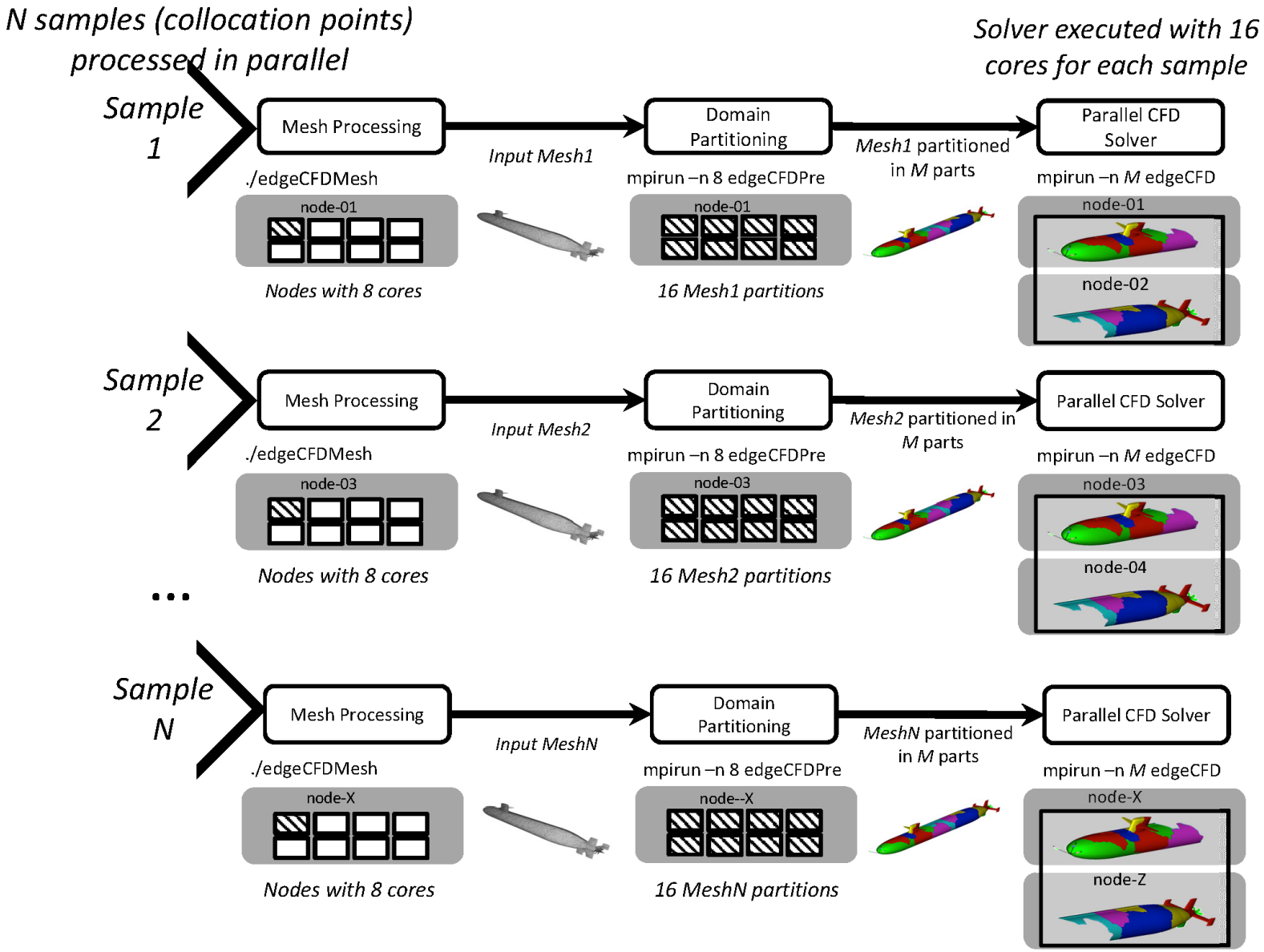


Turbulence UQ analysis



GUERRA, G. ; et al. **Analysis of flow-induced vibration model under uncertainties using an iterative workflow**. In: Int. Symposium on Uncertainty Quantification and Stochastic Modeling 2012

Two-level Parallel Strategy

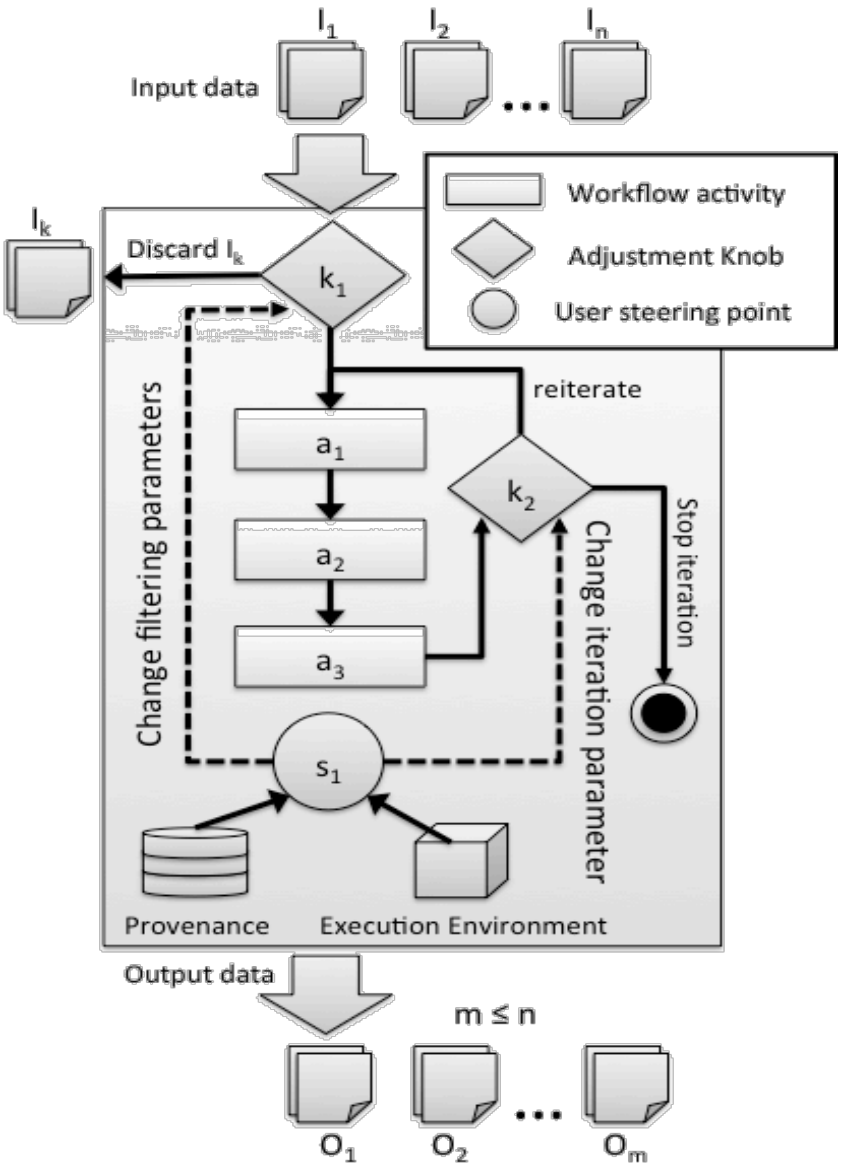


Adjusting the interpolation levels

- Execution restarts over and over
- UQ analyst may loose track of what has already been explored and how the UQ workflow evolved
- The user should be able to analyze partial results during execution
  - to dynamically interfere in the next steps of the workflow
  - instead of interrupting and resubmitting the workflow

Dias, J., et al. (2011) **Supporting Dynamic Parameter Sweep in Adaptive and User-Steered Workflow**. WORKS Workshop at IEEE Supercomputing

Dynamic Workflow example



- User steering points
  - Select provenance data
  - Trigger adjustments
  - Similar to checkpoints in Taverna
- Adjustable knobs
  - Store adjustable parameter
  - Change iteration
  - May affect the dataflow

Workflow execution
Off-line (black-box) X On-line (steering)

- Only after the whole workflow execution :
  - Check on data derivation & results
  - Change # interpolation levels
- Interrupt the execution

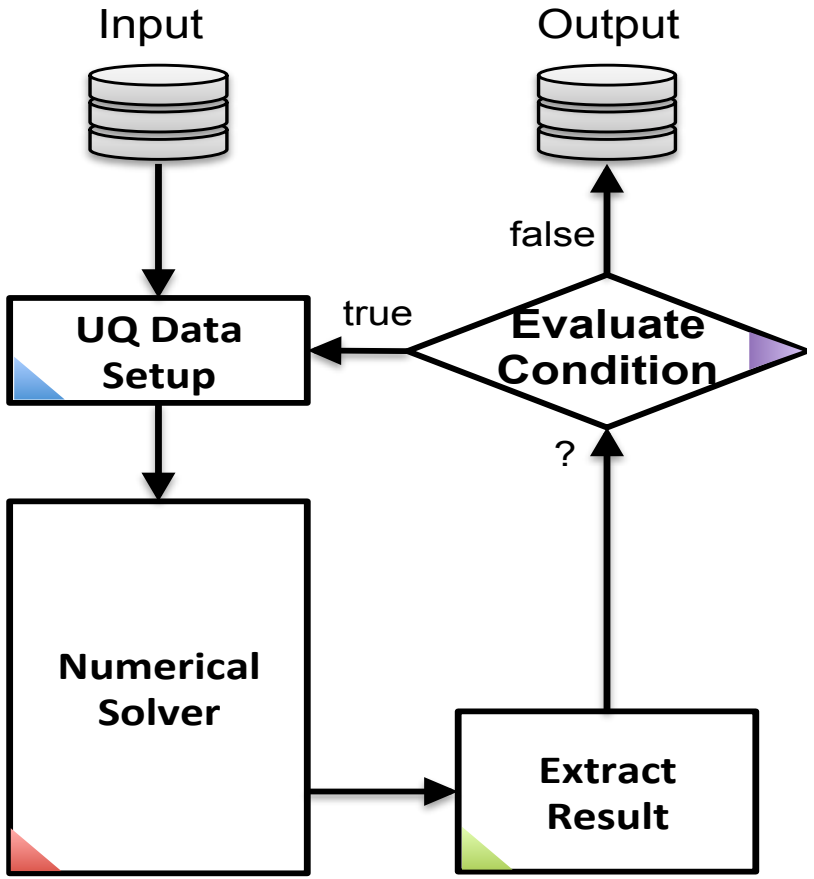
"Off-line"

During workflow execution

- Partial results & provenance are analyzed
- Snapshots of current simulation results to refine the model (iterations) during runtime
- **Fine tuning of parameters**
- **Interfere on loop specification**

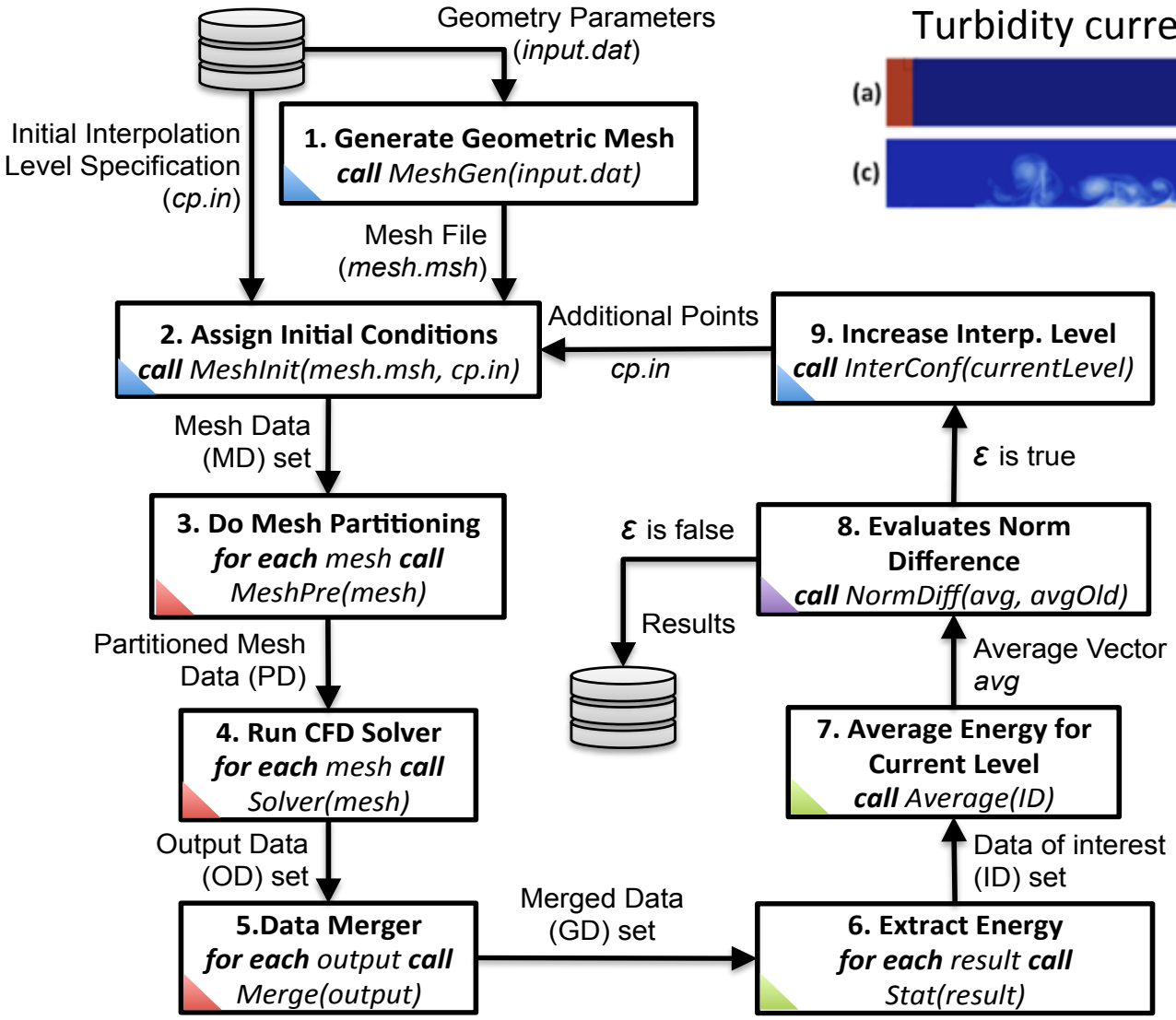
"On-line"

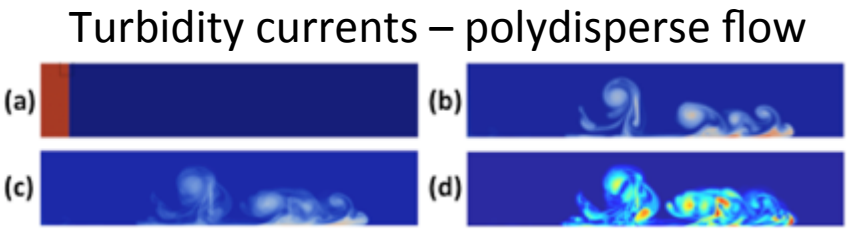
Executes for predefined levels with predefined condition



criterion is the difference between the vector norms and should be below a given **threshold**, initially defined as 0.001

UQ Executes for all predefined levels





Tuple generation of Activity 5
UQ parallel execution

R

ID	OD
1	out1.dat
2	out2.dat
...	
n	outN.dat

T ← *Map(Merge , R)*

```
#./Merge  -file out1.dat
```

DB Query

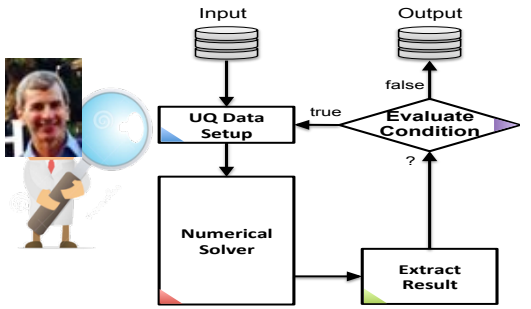
T

ID	Merged Data	Energy
1	GD1.dat	
2	GD2.dat	
...		
m	GDm.dat	

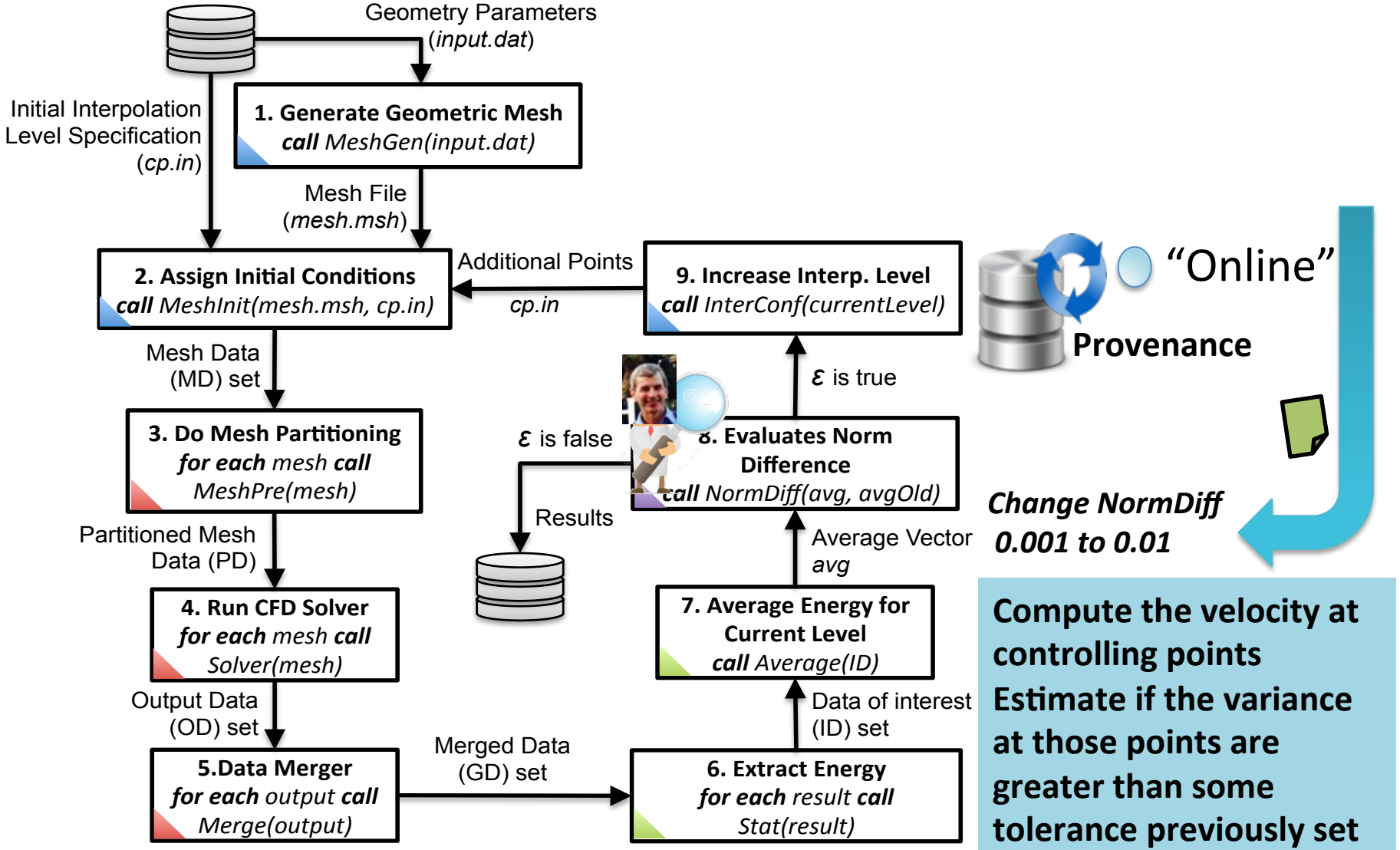
Energy
0.56
0.58
...
0.88

Results

Data Extractor

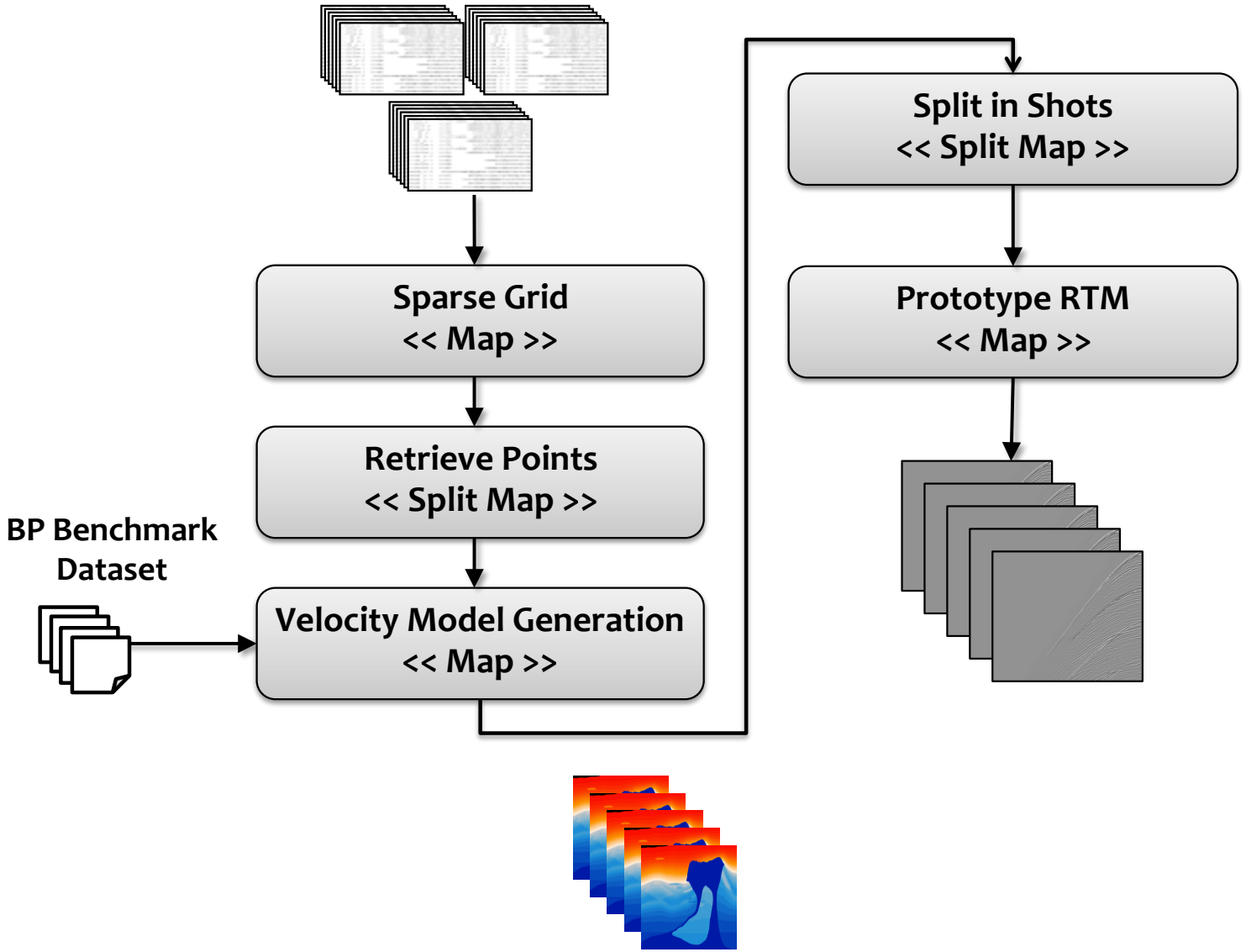


UQ With User Steering & Intervention

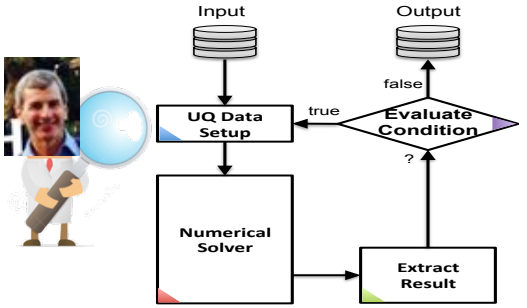


J. Dias, G. Guerra, F. Rochinha, A. Coutinho, P. Valduriez, M. Mattoso. Data-Centric Iteration in Dynamic Workflows. ***Future Generation Computer Systems***, Vol. 4, 114-126, 2015.

RTM + UQ Workflow



Concluding…



- Parallel Scientific Workflows & Provenance:
  - is aware of the dependencies and the data-flow
  - data-flow parallel execution
  - integration (indexing) of scientific resources (files)
  - big data analytics through provenance
- User steering & dynamic intervention
  - patterns that humans can easily detect but computer algorithms have a hard time finding
  - dynamic loops
- **Provenance DB act as an index to related raw data files**

Some Results: COPPE-UFRJ – INRIA

- Algebraic workflow engine (2012-2013)

- E. S. Ogasawara, J. Dias, V. Silva, F. S. Chirigati, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Chiron: a parallel engine for algebraic scientific workflows. ***Concurrency and Computation****: Practice and Experience*, 25(16):2327–2341, 2013.
- Chirigati, F S ; Sousa, V. ; Ogasawara, E. ; Oliveira, D. ; Dias, J. ; Porto, F. ; Valduriez, P. ; Mattoso, Marta . Evaluating Parameter Sweep Workflows in High Performance Computing. In**: Int Workshop on Scalable Workflow Enactment Engines and Technologies (SWEET'12**), 2012, Phoenix. SIGMOD.

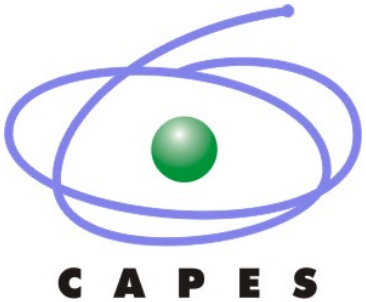
- Dynamic Loops in workflow execution (2013-2014)

- J. Dias, G. Guerra, F. Rochinha, A. Coutinho, P. Valduriez, M. Mattoso. Data-Centric Iteration in Dynamic Workflows. ***Future Generation Computer Systems***, Vol. 4, 114-126, 2015.
- J. Dias, E. S. Ogasawara, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Algebraic Dataflows for Big Data Analysis. **IEEE Bigdata Conference** 2013

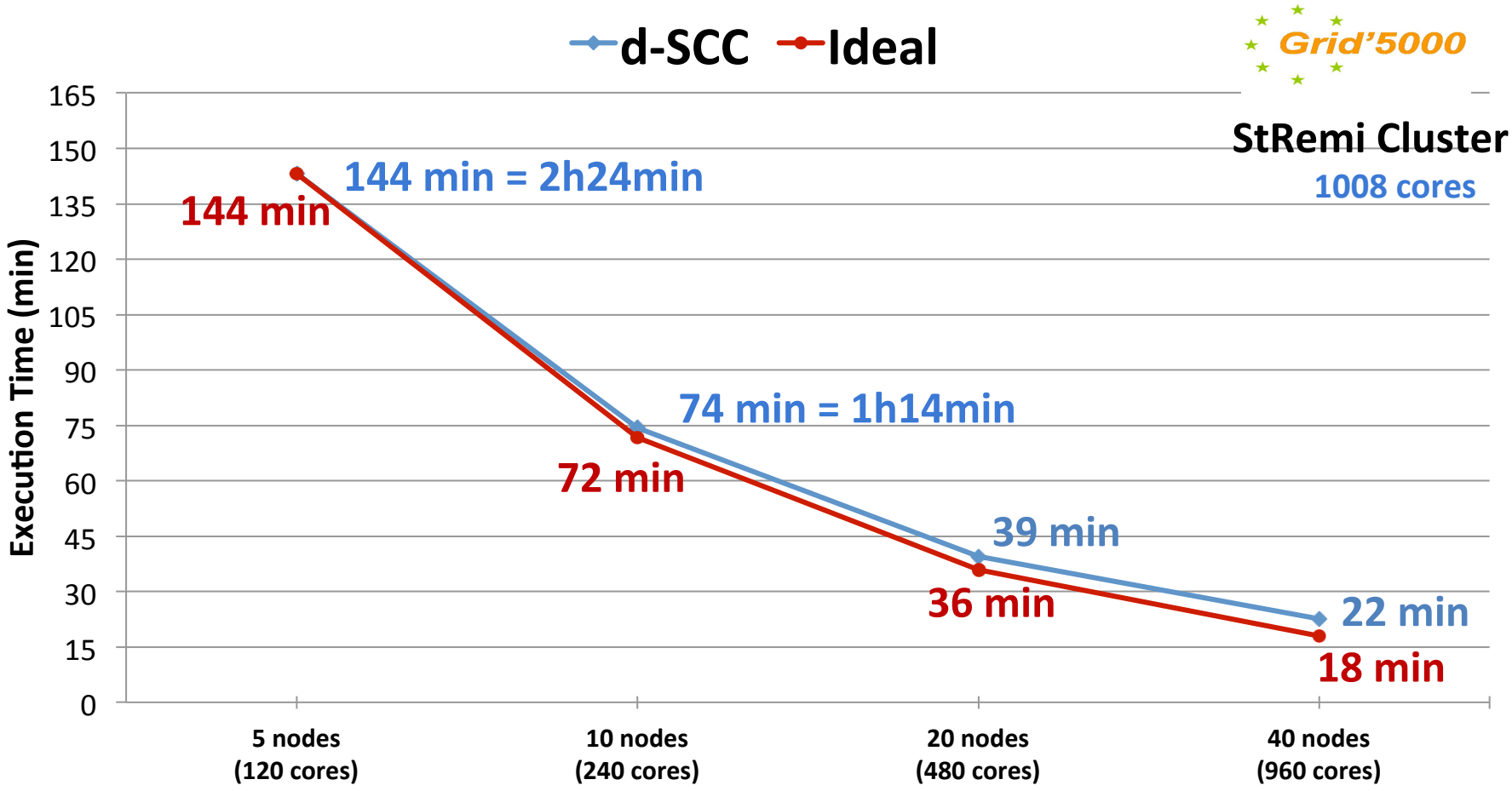
- Querying *in-situ* raw data files (2014-2015)

- V. Silva, D. de Oliveira, P. Valduriez, M. Mattoso. Analyzing Related Raw Data Files through Dataflows. ***Concurrency and Computation: Practice and Experience***, 2015
- R. Souza, V. Silva, D. de Oliveira, P. Valduriez, A. Lima, M. Mattoso. Parallel Execution of Workflows Driven by a Distributed Database Management System **IEEE/ACM SuperComputing Conference** *(SC15)*, 2015.

Acknowledgements



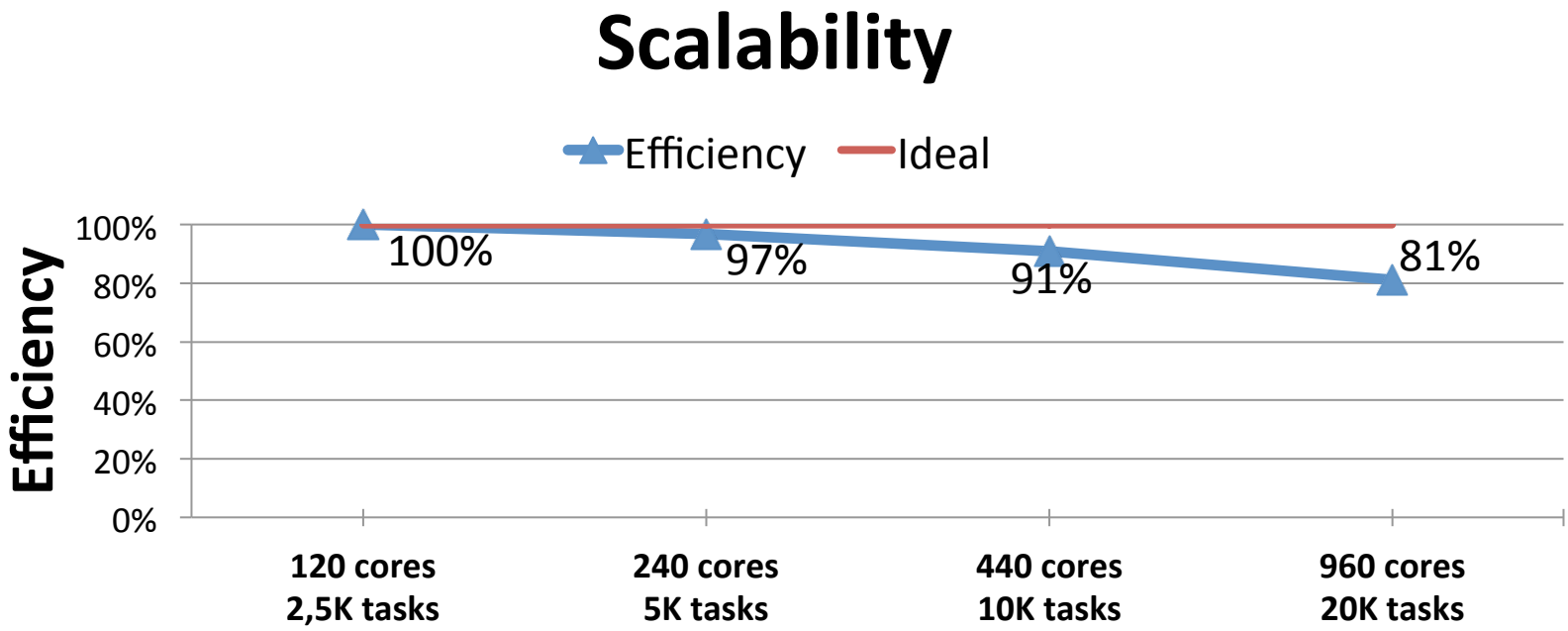
Workflow Execution Time close to ideal



30 k tasks
32 s tasks cost on average

Theoretical sequential time = 276 h = **11.5 days**

Scalability:
efficiencies over 80%



20k tasks
32 s tasks cost on average